Preuves traces d'éligibilité

On souhaite montrer que faire des mises à jour à chaque étape de la trajectoire selon :

$$V_\pi(s) \leftarrow V_\pi(s) + \alpha \delta_t e_t(s) \quad \forall s \in \mathcal{S}$$

avec

$$\delta_t = r_t + \gamma V_\pi(s_{t+1}) - V_\pi(s_t)$$

et $e_t(s)$ des traces d'éligibilité définies comme :

$$e_0(s) \leftarrow 0 \quad \forall s \in \mathcal{S}$$

$$e_t(s) \leftarrow \lambda\gamma e_{t-1}(s) + \mathbf{I}(S_t = s) \quad \forall s \in \mathcal{S}$$

est équivalent à faire une mise à jour globale en fin de trajectoire pour tout s_t de cette trajectoire :

$$V_\pi(s_t) \leftarrow V_\pi(s_t) + \alpha(G_t^\lambda - V_\pi(s_t))$$

Dans un premier temps on s'intéresse à la quantité $G_t^\lambda = \lim\limits_{T \to +\infty} (1-\lambda) \sum_{n=1}^T \lambda^{n-1} G_t^{(n)}$.

$$\begin{aligned}
(1-\lambda)\sum_{n=1}^{T} \lambda^{n-1} G_t^{(n)} &= (1-\lambda)\Big[r_t(1+\lambda+...+\lambda^{T-1}) \\
&\qquad + \gamma r_{t+1}(\lambda + \lambda^2 + ... + \lambda^{T-1}) \\
&\qquad + ... \\
&\qquad + \gamma^{T-1} r_{t+T-1}(\lambda^{T-1}) \\
&\qquad + \sum_{n=1}^{T} \gamma^n \lambda^{n-1} V_{t+n} \Big] \\
&= \sum_{n=1}^{T} \gamma^{n-1} r_{t+n-1}(\lambda^{n-1} - \lambda^T) + \sum_{n=1}^{T} \gamma^n V_{t+n}(\lambda^{n-1} - \lambda^n)
\end{aligned}$$

On a alors $G_t^\lambda = \sum\limits_{n=1}^{T} \gamma^{n-1}\lambda^{n-1} r_{t+n-1} + \sum\limits_{n=1}^{T} \gamma^n V_{t+n}(\lambda^{n-1} - \lambda^n)$ car $\lim\limits_{T \to +\infty} \lambda^T = 0$

En fin de trajectoire, pour tout état s rencontré, la somme des mises à jour effectuées pour cet état s selon les traces d'éligibilité correspond à :

$$V(s) \leftarrow V(s) + \alpha \sum_{t, st=s} \sum_{t'=t}^{T} (\lambda\gamma)^{t'-t} \delta_{t'}$$

Pour chaque s, pour tout t tel que $s_t = s$, on ajoute donc à $V_t = V(s_t)$ la quantité (pondérée par

α) suivante :

$$
\begin{aligned}
\sum_{t'=t}^{T}(\lambda\gamma)^{t'-t}\delta_{t'} &= r_t + \gamma V_{t+1} - V_t \\
&+ \lambda\gamma r_{t+1} + \lambda\gamma^2 V_{t+2} - \lambda\gamma V_{t+1} \\
&+ (\lambda\gamma)^2 r_{t+2} + \lambda^2\gamma^3 V_{t+3} - (\lambda\gamma)^2 V_{t+2} \\
&+ ... \\
&+ (\lambda\gamma)^{T-t-1} r_{T-1} + \lambda^{T-t-1}\gamma^{T-t} V_T - (\lambda\gamma)^{T-t-1} V_{T-1} \\
&+ (\lambda\gamma)^{T-t} r_T - (\lambda\gamma)^{T-t} V_T \\
&= \sum_{t'=t}^{T}(\lambda\gamma)^{t'-t} r_{t'} + \sum_{t'=t}^{T-1} \lambda^{t'-t}\gamma^{t'-t+1} V_{t'+1} - \sum_{t'=t}^{T}(\lambda\gamma)^{t'-t} V_{t'} \\
&= \sum_{n=1}^{T}(\lambda\gamma)^{n-1} r_{t+n-1} + \sum_{n=1}^{T} \lambda^{n-1}\gamma^{n} V_{t+n} - \sum_{n=1}^{T}(\lambda\gamma)^{n-1} V_{t+n-1}
\end{aligned}
$$

où la dernière égalité est obtenue en considérant nuls tout r_t et V_t pour $t > T$.
On note que $\sum_{n=1}^{T}(\lambda\gamma)^{n-1}V_{t+n-1} = \sum_{n=1}^{T}(\lambda\gamma)^{n}V_{t+n} + V_t$.
Il s'en suit alors que :

$$
\begin{aligned}
\sum_{t'=t}^{T}(\lambda\gamma)^{t'-t}\delta_{t'} &= \sum_{n=1}^{T}(\lambda\gamma)^{n-1} r_{t+n-1} + \sum_{n=1}^{T} \gamma^{n} V_{t+n}(\lambda^{n-1} - \lambda^{n}) - V_t \\
&= G_t^{\lambda} - V_t
\end{aligned}
$$